A Multi-Task Text Classification Pipeline with Natural Language Explanations for Greek Tweets

Nikolaos Mylonas[1], Nikolaos Stylianou[1], Theodora Tsikrika[1], Stefanos Vrochidis[1] and Ioannis Kompatsiaris[1]

Abstract

Interpretability has gained significant attention, with most such techniques producing rule-based or feature importance interpretations. While informative, these interpretations may be harder to understand for non-expert users and, therefore, cannot always be considered as adequate explanations. To that end, explanations in natural language are often preferred. This work introduces a novel pipeline for text classification tasks, offering predictions and explanations in natural language. It consists of (i) a classifier for providing the labels and (ii) an explanation generator to provide explanations. The proposed pipeline can be adopted by any text classification task, provided that ground truth rationales are available to train the explanation generator. Our experiments on sentiment analysis and offensive language identification in Greek tweets, use a Greek Large Language Model to obtain the necessary explanations that can act as rationales. The experimental evaluation, performed through a user study and based on three metrics, showed that this pipeline can produce adequate explanations when a sufficient amount of training data with accompanying explanations are available, even when these explanations are machine generated.

Keywords

Interpretability, Text classification, Textual explanations, CEUR-WS

1. Introduction

Machine Learning, particularly Deep Learning, is widely applied across domains where its predictions can influence critical processes, making interpretability essential for providing justifications. Interpretability comes in many forms, with the most common being rule based and feature importance interpretations [1]. These kinds of interpretations are not always preferred by non-expert users, as they may lack information in case of the former, or are not as intuitive in case of the latter. To that end, explanations in natural language, are becoming increasingly popular, as they are more easily understood by end users, while also containing the necessary information to explain the outcomes of machine/deep learning models.

This work explores the concept of multi-task pipelines that can provide both predictions and also explanations in natural language. This concept of multi-task predictions, containing both the labels and the corresponding explanations, has been explored mostly for sequence-to-sequence models [2], in which a single sequence generation model performs both tasks. In our

4th International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2025, June 19-20, 2025, Thessaloniki, Greece.

mylonik@iti.gr (N. Mylonas); nstylia@iti.gr (N. Stylianou); theodora.tsikrika@iti.gr (T. Tsikrika); stefanos@iti.gr (S. Vrochidis); ikom@iti.gr (I. Kompatsiaris)
0000-0002-5733-543X (N. Mylonas); 0000-0002-6396-5374 (N. Stylianou); 0000-0003-4148-9028 (T. Tsikrika); 0000-0002-2505-9178 (S. Vrochidis); 0000-0001-6447-9020 (I. Kompatsiaris)

work, we propose a pipeline that first provides predictions for text classification problems and then combines the predictions with their associated input to provide explanations in natural language for the predicted label, through the use of a sequence-to-sequence model. Unlike a single multi-task model that generates both labels and explanations simultaneously, the proposed pipeline allows for greater versatility by enabling the use of distinct models for each task, with independent performance, facilitating easier optimization.

We evaluate our pipeline for two text classification problems, sentiment analysis and offensive language detection, in the context of a low resource language, namely Greek, on two datasets originating from X (formerly Twitter) posts [3, 4] annotated with relevant sentiments and labels, respectively. To obtain the rationales needed to train the sequence-to-sequence model, we make use of a Greek Large Language Model (LLM) to generate explanations for each instance of the training/validation sets through prompting. Due to the lack of gold standard explanations for both datasets, the performance of our pipeline is evaluated through a user study on the explanations produced by the sequence-to-sequence model on the test set using three different explainability metrics: *Plausibility* [5], *Coherence* [6], as well as a new metric introduced in this work, referred to as *Perfidiousness*. We make our code and user study results publicly available[1].

2. Related Work

Regarding sentiment analysis in Greek, notable works include lexicon- and aspect-based analysis of tweets [7, 8], and the development of annotated Greek datasets [3]. For offensive language identification in Greek, a study introduced a manually annotated Greek tweets dataset [4].

Regarding interpretability, techniques like LIME [9], Integrated Gradients [10], and SHAP [11] assign weights to features based on their contribution to the output. While informative, these feature importance interpretations can be difficult for non-experts to understand. To address this, recent works focus on generating textual explanations [12].

When human-annotated rationales are available, metrics like Simulatability [13] evaluate explanations by testing if one model can predict another's outputs using the explanations. Unsupervised metrics such as Robustness [14], Comprehensibility [15], Faithfulness [16], and Plausibility [5] assess stability, informativeness, alignment with predictions, and human persuasiveness. User studies are ideal for evaluating explanations, but may introduce bias [17, 18].

Self-rationalising models [19] provide both predictions and explanations, often using sequence-to-sequence models trained with ground truth labels and rationales [2, 20]. These models generate free-text explanations for tasks like natural language inference [21] and machine translation [22]. Alternatively, pipelines can handle combined tasks, where one model makes predictions, and another generates rationales based on the input and prediction [19].

3. Text Classification with Natural Language Explanations

This work introduces a pipeline for text classification that generates both predictions and natural language explanations through a two-step process: a classification model predicts the label of a

[1] Released upon publication

given text, and a sequence-to-sequence model generates explanations by combining the input text with the predicted label using a conditional generation approach. Our proposed pipeline is very versatile, making it applicable to any text classification task, provided that ground truth rationales are available to fine-tune the explanation model.

The two models are trained independently and used sequentially during inference. The first model (*classifier*) handles textual input and produces predictions. The second model (*explanation generator*) provides natural language explanations by incorporating both the input text and its predicted label into a composite text format (e.g., "*input text* has *label* label"), ensuring explanations support the predicted label. Ground truth rationales for each training instance are required in order to train this model. Training datasets are preferably aligned, but can also vary. Once trained, the pipeline predicts labels for new texts and generates explanations, enhancing the transparency of Machine Learning models, while serving as a robust classification tool.

3.1. Datasets and Model Selection

We used two datasets, the first [3] includes politically themed tweets annotated with *Positive*, *Negative*, or *Neutral* sentiments; this dataset is highly imbalanced, with only 4.83% of the 1640 instances labelled as *Positive*. The second [4] categorises tweets as *Offensive* or *Not Offensive*, and is also imbalanced, with 28.52% of the 3345 instances labelled as *Offensive*. Both datasets were split into training/validation/test sets using a 70%/10%/20% scheme.

To create the rationales for training the *explanation generator*, we used the Greek LLM *Meltemi* [23], built on *Mistral-7B* [24] and trained on a large corpus of high-quality Greek texts. Using its instruction-tuned variant, *Meltemi-7B-Instruct-v1*, we designed a custom sequence of prompts to obtain explanations, addressing the absence of ground truth rationales.

We used two prompts (Table 1), one to set the desired output format and another to query the model with input text and its label, requesting an explanation. Queries were in Greek, with translations provided for clarity. Originally designed for sentiment analysis, the prompts were

Table 1
Greek and Translated Prompts used for the Greek LLM Meltemi, along with a generated explanation

Greek	**Translation**
Conditioning Prompt Θα σου δώσω ένα κείμενο το οποίο έχει χαρακτηριστεί με ένα sentiment. Θέλω να μου επιστρέψεις μια πρόταση μόνο που να επεξηγεί τον λόγο για τον οποίο το κείμενο αυτό να χαρακτηριστεί με το sentiment αυτό. Μην γράψεις τίποτα άλλο πέρα από την πρόταση που να επεξηγεί το sentiment.	**Conditioning Prompt Translation** I will give you a text that has been labeled with a sentiment. I want you to return only one sentence explaining why this text has been labeled with this sentiment. Do not write anything other than the sentence explaining the sentiment.
Query Prompt Το κείμενο: {input text} έχει χαρακτηριστεί με το ακόλουθο sentiment {label}. Γράψε μου μια πρόταση που να εξηγεί γιατι το κείμενο χαρακτηρίστηκε με το sentiment.	**Query Prompt Translation** The text: {input text} is labelled with the following sentiment {label}. Write a sentence explaining why the text is labeled with the sentiment.

minimally adapted for offensive language identification. These prompts generated explanations for training and validation instances, which were used to train the explanation generator.

For our experiments, we used Greek-BERT [25] as the *classifier*, leveraging its pre-training on a large Greek corpus. The model was fine-tuned for 15 epochs on the sentiment analysis dataset and for 10 epochs on the offensive language identification dataset. For the *explanation generator*, we selected BART [26], a versatile sequence-to-sequence model known for tasks like summarisation and translation. BART was fine-tuned for 15 epochs on both datasets.

3.2. User-Centred Evaluation

The proposed pipeline, comprising a *classifier* and an *explanation generator*, is evaluated separately for each component. The classifier's performance on sentiment analysis and offensive language identification tasks, both of which have gold-standard annotations, is assessed using F1-Score and Balanced Accuracy. Since no ground truth rationales exist for the generated explanations, we could not use metrics like Simulatability [13], instead we focused on a user-centred study to evaluate the explanation generator.

We evaluated the quality of the generated explanations through a user study using three metrics, including *Plausibility* and *Coherence*. The *Plausibility* metric assesses how convincing an explanation is in justifying the predicted label on the input text, regardless of whether the label itself is correct. The *Coherence* metric evaluates how well-formed the explanation is, reflecting its similarity to human-written text and absence of grammatical or syntactical errors, with low coherence indicating a lack of meaningful structure.

We also propose a novel metric, *Perfidiousness*, to evaluate how effectively a generated explanation represents a label other than the predicted one. High scores indicate that the explanation faithfully supports an alternative label, while low Perfidiousness reflects alignment with the predicted label, highlighting the explanation generator's ability to capture label-specific information from the input text. For example, if a Neutral sentiment prediction is accompanied by an explanation justifying Neutral sentiment, Perfidiousness would be low; however, if the explanation argues for a Positive or Negative sentiment, Perfidiousness would be high.

Two user studies were conducted, one for each dataset, with 15 native Greek-speaking participants who all had a technical background and some familiarity with machine/deep learning, but not necessarily with explainability. For each study, participants evaluated 10 random instances per label, resulting in 30 instances for the sentiment analysis dataset and 20 for the offensive language identification dataset. Due to a limited number of positive sentiment examples in the sentiment analysis dataset, we selected 8 positive instances, 11 neutral, and 11 negative to maintain balance. No such issue arose for the offensive language dataset. The same instances were presented to all users, who were provided with the input text, predicted label, and explanation. Participants rated explanations on a scale from 1 to 10 for each metric, with final scores calculated as the average rating per instance and then averaged across all instances.

4. Experimental Results

Our experimental results focus primarily on evaluating the generated textual explanations, rather than the classification task itself, as both studied problems are generally well-solved with

Table 2
Average performance of explainability metrics per examined sentiment (top) and label (bottom)

Sentiment	Plausibility	Coherence	Perfidiousness
Neutral	9.00	7.68	1.41
Negative	6.34	5.06	2.41
Positive	5.46	4.29	2.96
Overall	7.08	5.82	2.19

Label	Plausibility	Coherence	Perfidiousness
Offensive	7.93	7.50	1.83
Not Offensive	8.45	7.37	1.67
Overall	8.19	7.44	1.75

proper tuning. For sentiment analysis, the classifier achieved 92.9% Balanced Accuracy, 79.8% macro F1-Score, and per-sentiment F1 scores of 58.3% $F1_{Pos}$, 87.5% $F1_{Neg}$, and 93.6% $F1_{Neu}$. The model performed significantly better for the Neutral and Negative sentiments due to the larger availability of training data, while performance dropped for the Positive sentiment because of limited examples. Regarding the explanations, results (Table 2 top) reveal high Plausibility and average Coherence, with low Perfidiousness, indicating that most explanations align with the predicted sentiment. The imbalance in training data seems to impact the quality of explanations, with Neutral sentiment explanations exhibiting higher Plausibility and Coherence, followed by Negative and then Positive sentiments.

For offensive language identification, the classifier achieved 86.6% Balanced Accuracy, 85.5% macro F1, and per-label F1 scores of 91.0% $F1_{NotOff}$ and 80.0% $F1_{Off}$. Unlike the sentiment analysis dataset, this dataset's balanced nature led to consistent classifier performance across labels. Explanations for this dataset (Table 2 bottom) scored higher in Plausibility and Coherence, likely due to sufficient training examples for both labels. The lower Perfidiousness further suggests that the explanation generator produces more faithful explanations when adequate training data are available, as it can better distinguish between labels and align explanations with the predicted output.

5. Conclusions

In this work, we introduced a pipeline combining two independent models a classifier for predictions and a sequence-to-sequence model for generating explanations. Experiments demonstrated that the pipeline produces explanations that are generally coherent and informative for users, while maintaining the classifier's performance on the primary text classification task. Furthermore, the quality of explanations improves with more training data, enabling the explanation generator to produce explanations that are both more plausible and coherent.

As future work, we aim to expand our experiments to include more datasets, particularly in English, ideally with human-annotated rationales, as increased data availability has shown potential to improve explanation quality in both Plausibility and Coherence. A larger, more diverse user study will also be conducted to obtain a diversified user sample for evaluation. Additionally, we aim to explore using a single self-rationalising model for simultaneous predictions and explanations, comparing its performance to our pipeline. Testing alternative models for both the classifier and explanation generator is another direction for further research.

Acknowledgments

This work has received funding by the European Union Horizon – Research and Innovation Framework Programme, under grant agreement No 101121282. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] M. Saarela, S. Jauhiainen, Comparison of feature importance measures as explanations for classification models, SN Applied Sciences 3 (2021) 272. URL: https://doi.org/10.1007/s42452-021-04148-9. doi:10.1007/s42452-021-04148-9.

[2] H. Liu, Q. Yin, W. Y. Wang, Towards explainable nlp: A generative explanation framework for text classification, 2019. arXiv:1811.00196.

[3] A. Tsakalidis, S. Papadopoulos, R. Voskaki, K. Ioannidou, C. Boididou, A. I. Cristea, M. Liakata, Y. Kompatsiaris, Building and evaluating resources for sentiment analysis in the greek language, Language Resources and Evaluation 52 (2018) 1021–1044.

[4] Z. Pitenis, M. Zampieri, T. Ranasinghe, Offensive language identification in Greek, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 5113–5119. URL: https://aclanthology.org/2020.lrec-1.629.

[5] B. Herman, The promise and peril of human evaluation for model interpretability, 2019. arXiv:1711.07414.

[6] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3583558. doi:10.1145/3583558.

[7] D. Kydros, M. Argyropoulou, V. Vrana, A content and sentiment analysis of greek tweets during the pandemic, Sustainability 13 (2021). URL: https://www.mdpi.com/2071-1050/13/11/6150. doi:10.3390/su13116150.

[8] G. Aivatoglou, A. Fytili, G. Arampatzis, D. Zaikis, N. Stylianou, I. Vlahavas, End-to-end aspect extraction and aspect-based sentiment analysis framework for low-resource lan-

guages, in: K. Arai (Ed.), Intelligent Systems and Applications, Springer Nature Switzerland, Cham, 2024, pp. 841–858.

[9] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 (2016). URL: https://arxiv.org/abs/1602.04938.

[10] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, 2017. arXiv:1703.01365.

[11] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[12] P. Jha, K. Maity, R. Jain, A. Verma, S. Saha, P. Bhattacharyya, Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations, 2024. arXiv:2401.09899.

[13] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017. arXiv:1702.08608.

[14] D. A. Melis, T. Jaakkola, Towards robust interpretability with self-explaining neural networks, in: Advances in Neural Information Processing Systems, Montreal, Canada, 2018, pp. 7775–7784.

[15] M. Robnik-Sikonja, M. Bohanec, Perturbation-based explanations of prediction models, in: Human and Machine Learning - Visible, Explainable, Trustworthy and Transparent, Springer, International, 2018, pp. 159–175. doi:10.1007/978-3-319-90403-0_9.

[16] C. S. Chan, H. Kong, L. Guanqing, A comparative study of faithfulness metrics for model interpretability methods, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5029–5038. URL: https://aclanthology.org/2022.acl-long.345. doi:10.18653/v1/2022.acl-long.345.

[17] P. Lertvittayakumjorn, F. Toni, Human-grounded evaluations of explanation methods for text classification, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, November 3-7, ACL, Hong Kong, China, 2019, pp. 5194–5204. doi:10.18653/v1/D19-1523.

[18] B. Herman, The promise and peril of human evaluation for model interpretability, ArXiv abs/1711.07414 (2017).

[19] S. Wiegreffe, A. Marasović, N. A. Smith, Measuring association between labels and free-text rationales, 2022. arXiv:2010.12762.

[20] Z. Tang, G. Hahn-Powell, M. Surdeanu, Exploring interpretability in event extraction: Multitask learning of a neural event classifier and an explanation decoder, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Online, 2020, pp. 169–175. URL: https://aclanthology.org/2020.acl-srw.23. doi:10.18653/v1/2020.acl-srw.23.

[21] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom, e-snli: Natural language

inference with natural language explanations, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 31, Curran Associates, Inc., 2018.

[22] U. Ehsan, B. Harrison, L. Chan, M. O. Riedl, Rationalization: A neural machine translation approach to generating natural language explanations, 2017. arXiv:1702.07826.

[23] L. Voukoutis, D. Roussis, G. Paraskevopoulos, S. Sofianopoulos, P. Prokopidis, V. Papavasileiou, A. Katsamanis, S. Piperidis, V. Katsouros, Meltemi: The first open large language model for greek, 2024. URL: https://arxiv.org/abs/2407.20743. arXiv:2407.20743.

[24] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. arXiv:2310.06825.

[25] J. Koutsikakis, I. Chalkidis, P. Malakasiotis, I. Androutsopoulos, Greek-bert: The greeks visiting sesame street, in: 11th Hellenic Conference on Artificial Intelligence, SETN 2020, Association for Computing Machinery, New York, NY, USA, 2020, p. 110–117. URL: https://doi.org/10.1145/3411408.3411440. doi:10.1145/3411408.3411440.

[26] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. arXiv:1910.13461.